DOCUMENT RESUME

ED 449 222 TM 032 367

AUTHOR	Homack, Susan R.
TITLE	Understanding What ANOVA Post Hoc Tests Are, Really.
PUB DATE	2001-02-01
NOTE	15p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Analysis of Variance
IDENTIFIERS	*Post Hoc Tests; Scheffes Contrast Test; Tukeys Test

ABSTRACT

Many people use post hoc tests but do not completely understand why regular t-tests are not used post hoc or exactly what these tests are doing. This paper makes a direct comparison of Tukey and Scheffe post hoc tests with regular t-tests conducted using a new testwise alpha to make clear that controlling experimentwise error through some kind of Bonferroni-type correction is the basic concept underlying post hoc methods. (Contains 11 references.) (Author/SLD)

Running head: UNDERSTANDING ANOVA POST HOC TESTS

Understanding What ANOVA Post Hoc Tests Are, Really

Texas A & M University 77843-4225

Susan R. Homack

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1, 2001

Abstract

Many people use post hoc tests, but do not completely understand why regular t-tests are not used post hoc, or exactly what these tests are doing. The paper will make a direct comparison of Tukey and Scheffe post hoc tests with regular t-tests conducted using a new testwise alpha (α_{tw}*) to make clear that controlling experimentwise error via some kind of Bonferroni-type correction is the basic concept underlying post hoc methods.

Understanding What ANOVA Post Hoc Tests Are, Really

Many people use post hoc tests but do not completely understand why conventional t-tests are not used post hoc. The paper provides a description of the t-test, a brief explanation and history of ANOVA, and a direct comparison of both simple (e.g., Tukey) and complex (e.g., Scheffe) post hoc tests with regular t-tests.

T-tests and "Experimentwise" Error

Prior to the 1920's, when a researcher tested the hypothesis that K sample means were equal, where k>2, the procedure used was the t-test. The formula for t-test with two independent samples is:

$$t= \frac{\overline{X}_1 - \overline{X}_2}{s_{x1-x2}}$$

where s_x= the sample distribution standard deviation of the mean differences . When conducting experiments with numerous sample means, many separate t-tests (k*(k-1)/2) must be conducted in order to compare all possible pairs of means.

The number of t-tests that must be computed is the number of pairs of means that can be contrasted. If no "complex" combinations of means (such as the mean of groups 1 and 2 combined versus the mean of group 3) are being conducted, the number of "simple" pairs of means equals [k (k –1)] / 2. For example, if 4 means were being compared, the number of pairwise t-tests would be:

$$[\,4\,(4-1)]\,/\,2$$
$$[4\,(3)\,]\,/\,2$$
$$12\,/\,2 = 6.$$

Specifically, the 6 "simple" contrasts that could be tested with t-tests would be:

H_o: $M_1 = M_2$
H_o: $M_1 = M_3$
H_o: $M_1 = M_4$
H_o: $M_2 = M_3$
H_o: $M_2 = M_4$
H_o: $M_3 = M_4$

Although it might appear that using multiple t-tests is appropriate, there are problems with computing too many t-tests. When more than one t-test is computed, each at a specific level of significance (such as α=.05), the probability of making one or more "experimentwise" Type I errors in the series of t-test is greater than α. "Experimentwise" error (α_{ew}) refers to the probability of making one or more Type I errors (rejecting a true null hypothesis) anywhere in the full set of all hypothesis tests, where these tests are each conducted at a given "testwise" alpha level (e.g., α_{tw}=.05).

The "experimentwise" error rate can be readily computed (Love, 1988) whenever either the correlations of the dependent variables or of the hypotheses are all either (a) 1.0 or (b) 0.0. If these correlations are all 1.0, then $\alpha_{ew}=\alpha_{tw}$. It is only when there is a perfect correlation or only one hypothesis is tested that there will be no inflation of EW error rate, because in actuality still only one hypothesis is being tested when all variables are perfectly correlated (Thompson, 1994a). If the hypotheses are at all uncorrelated, then there will be at least some inflation of the experimentwise error probability when several hypotheses are tested. If the scores or hypotheses are perfectly uncorrelated (e.g., as in the factorial multiway ANOVA –Thompson, 1994a), then according to the Bonferroni formula:

α_{ew}= 1- $(1-\alpha_{tw})^k$,

where k is the number of uncorrelated scores or hypotheses. The inflation is at its maximum when the hypotheses are perfectly uncorrelated.

Researchers recognize that "When multiple statistical tests are carried out in inferential data analysis, there is a potential problem of 'probability pyramiding '" (Huberty & Morris, 1989, p. 306). For example, if 10 uncorrelated tests are conducted at the α_{tw}=.05 level of statistical significance, the "experimentwise" Type I error rate according to the Bonferroni formula is inflated to :

$$1 - (1-0.05)^{10} = 0.40126.$$

One would know that it is more than 40% likely that a Type I error was being made in the study. As Huberty and Morris (1989 p. 306) indicated, "Use of conventional levels of Type I error probabilities for each test in a series of statistical tests may yield an unacceptably high Type I error probability across all the tests (the "experimentwise error rate")."

Witte (1985) provides an analogy that may further clarify probability pyramiding:

> When a fair coin is tossed only once, the probability of heads equals 0.50—just as when a single t-test is to be conducted at the 0.05 level of significance, the probability of a type 1 error equals 0.05. When a fair coin is tossed three times, heads can appear not only on the first toss but also on the second or third toss, and hence the probability of heads on *at least one* of the three tosses exceeds 0.05. By this same token, when a type I error can be committed not only on the first test but also on the second or third test, and hence the probability of committing a Type I error on *at least one* of the three tests exceeds 0.05. In fact the cumulative probability of at least one Type I error can be as large as 0.15 for this series of three t-tests. (p. 236)

This coin flip example illustrates a worst-case inflation of EW error (analogized as the flip of a head—H), because the results of each flip are perfectly uncorrelated with previous results (the coin presumably being unaware of or unaffected by its previous behavior.)

Thompson (1994a) further explains the EW error rate in studies that use t-tests:

> The EW error rate in a study ranges somewhere between the nominal TW alpha level (when only one test is conducted or all hypotheses are perfectly correlated) and [1-(1-testwise alpha)] raised to the power of the number of hypotheses tested (when more than one test is conducted and the hypotheses are perfectly uncorrelated. (pp. 6-7)

Love (1988) presented an example of the formula for estimating maximum inflation of EW Type I error, "As an example involving estimation of EW error rate, if nine hypotheses were each tested at the 0.05 level in a single study, the experimentwise error rate would range somewhere between .05 and .37." She also presented the Proof of the Bonferroni formula.

Furthermore, although one would know that if there is "experimentwise" Type I error inflation in the previous examples, one would not know (a) how many Type I errors were being made, or (b) which of the statistically significant results were the Type I errors (Thompson, 1994b).

Fisher's ANOVA

In light of the potential problems of 'probability pyramiding' (Huberty & Morris, 1989) that are present in t-tests, Fisher's articulation of analysis of variance (ANOVA) in the 1920's was extremely important. Following the invention of the ANOVA, researchers who wanted to compare means of more than k=2 groups no longer had to conduct [K (k-1)/2] t-tests for all the pairwise combinations of group means.

In other words, the use of a one-way ANOVA rather than a series of pairwise t-tests meant that "experimentwise" Type I error rates were no longer inflated by conducting a series of t-tests. Instead, a single one-way omnibus F-test could be used to test for differences among the set of k means while maintaining the Type I error rate at the pre-established alpha level for the entire set of comparisons. As Hinkle, Wiersma, and Jurs (1998 p.351) indicated, "In a one-way ANOVA, the total variance can be divided into two sources: (1) variation of scores *within* groups and (2) variation *between* the group means and the grand mean."

Although Fisher developed the ANOVA in the 1920's, it was not until the 1960's when educational research training expanded that the use of the ANOVA became prominent (Willson, 1980). Edington (1974) tabulated the inferential statistical procedures used in seven American Psychological Association (APA) journals from 1948 to 1972. Over this 25-year period, the use of the ANOVA increased dramatically from 11% of the 1948 articles to 71% of the 1972 articles. The use of t-tests decreased concomitantly, from 51% in 1948 to 12% in 1972. More recently, Elmore and Woehlke (1988) reviewed literature published in the American Educational Research Journal (AERJ), Educational Researcher (ER), and the Review of Educational Research (RER) from 1978 to 1987. Again, the most frequent statistical method to appear in the journal was the analysis of variance.

Post Hoc Tests

In classical ANOVA, if (a) statistically significant omnibus effects are obtained and (b) the number of k groups is greater than 2, then "post hoc" tests must be conducted to determine which groups differ. For example, with three groups, it might be that two means are equal but the third is higher, or maybe two means are equal but the third is lower, or it could be that all three means are different. By using a post hoc procedure, the researcher attempts to probe the data to find out which of the possible non-null scenerios is most likely to be true. Various terms are used in a synonymous fashion to mean the same thing as the term 'post hoc test.' The three synonyms that show up most often in the published data are "*a posteriori* test", "follow-up test", and "unplanned comparison test" (Huck, 2000).

Post hoc multiple comparison tests, unlike the t-test, were developed to maintain the *a priori* Type I error rate when computing a series of comparisons following the rejection of the null hypothesis in the ANOVA. A post hoc test can be defined as *a t-test with a built in Bonferroni-type correction that takes into account all sample means within a study.* In a post hoc test, the researcher's goal is to better understand why the ANOVA H_0 was rejected. Because the H_0 indicated that equality exists among all population means, we can say that a set of post hoc comparisons is designed to help the researcher gain insight into the pattern of means. Huck (2000) discussed how research hypotheses . drive the use of post hoc investigations:

> It should not be surprising that differences in research hypotheses lead researchers to do different things in their post hoc investigations. Sometimes for example, researchers set up their post hoc investigations to compare each sample mean against

> every other sample mean. On other occasions they use their post hoc test to compare the means associated with each of several experimental groups against a control group's mean, with no comparisons made among the experimental groups. On rare occasions, a post hoc investigation is implemented to compare the mean of one of the comparison groups against the average of the means of two or more of the remaining groups. (pp. 356-357)

"Simple" Post Hoc Tests: Tukey Method

So-called "simple" post hoc tests evaluate whether the means of two groups are the same (e.g., $\underline{M}_{k=1} = \underline{M}_{k=3}$). There are several formulas for doing simple post hoc tests. For the purpose of this paper, the Tukey post hoc test will be discussed. The Tukey test is one of the more conservative simple post hoc tests that exerts considerable control over Type I errors (Huck, 2000). The Tukey method, often called the HSD (honestly significant difference) test, is designed to make all pairwise or simple comparisons while maintaining the experimentwise error rate (α_{ew}) at the pre-established α level (Hinkle, et al., 1998). The null hypothesis tested for each pairwise comparison is

$$H_o: X_a = X_b \text{ for } a \neq b$$

That is, each pair of population means is equal. The test statistic is Q defined as follows:

$$Q = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{MS_w/n}}$$

The Tukey post hoc test uses the studentized range (Q) distributions to maintain the experimentwise alpha at the a priori alpha level (Hinkle et al., 1998). The Q distributions were developed to determine the minimum difference between the largest and the smallest means in a set of K sample means that is necessary to reject the hypothesis that the corresponding population means is equal.

Post Hoc Test for "Complex" Comparisons: Scheffe Method

When a researcher is interested in testing hypotheses that are more complex than simple differences between pairs of means, a complex comparison is made. For example, one might want to know whether two experimental groups (considered together) differ from a control group. The Scheffe method is the most versatile and most conservative procedure that can be used to test these complex hypotheses. With the Scheffe method, the form of the null hypothesis is as follows:

H_o: $\Sigma C_k \mu_k = 0$

where $\Sigma C_k = 0$.

That is, we add (or subtract) the products of the means multiplied by these coefficients. The only restriction is that, for each hypothesis, the sum of the coefficients must equal zero (Hinkle et al., 1998). For any contrast, the coefficients (C_k) are nonzero only for the population means under consideration in the hypothesis. For any population mean not included in the hypothesis, $C_k = 0$.

Suppose a researcher had four samples of people in a study and was interested in determining whether people in group two differed on a chosen variable from the three other samples. This hypothesis can be written as

$$H_o: \mu_2 = \frac{\mu_1 + \mu_3 + \mu_4}{3}$$

This hypothesis can be expressed in the terms of a contrast:

$$H_0: (1/3)\mu_1 + (-1)\mu + (1/3)\mu_3 + (1/3)\mu_4 = 0$$

Notice that the sum of the coefficients equals to zero. When there are unequal group sizes, it is necessary to adjust the coefficients. Details of the required procedure are available (Hinkle et al.,1998). The test statistic for the Scheffe method is:

$$F = \frac{(\Sigma C_k X_k)^2}{(MS_w)\,[\Sigma[C_k^2/n_k)]}$$

The critical value is determined by multiplying the critical value of F used in the ANOVA by the factor K-1, where K is the number of groups (Hinkle et al., 1998). This multiplication results in an increase in the critical value and is the primary reason why the Scheffe Method is a conservative method.

Comparisons of t-tests with Post Hoc Test

Post hoc tests are conceptually related to conducting regular t-tests, but where the t-tests are conducted with an "adjusted" testwise alpha (α_{tw}). Essentially the post hoc tests are regular t-tests, except that they build in a correction for experimentwise error, by lowering the α_{tw} to equal α_{tw}*, where this new testwise α is computed with some approximation of:

$$\alpha_{tw}* = \alpha_{tw} / j,$$

where j is the number of post hoc tests being conducted. By dividing the old α_{tw} by the number of post hoc tests conducted, every sample is taken into account (Huck, 2000).

Likewise, in the Tukey and Scheffe formulas, the MS_w in the denominator takes into account every observation in every sample. MS_w is a variance estimate. The within groups variance estimates are found by dividing the sum of squares within (SS_w) by the degrees of freedom associated with each of these estimates (Jurs, 1998). Sum of squares within is the sum of squares within each group, pooled across groups and then divided by the sum of degrees of freedom for each group.

Instead of dealing with the problem of an inflated Type I error risk by adjusting the level of significance as is done with the Bonferroni technique, the Tukey and Scheffe procedures make an adjustment in the size of the critical value used to determine whether an observed difference between two means is statistically significant. To compensate for the fact that more than one comparison is made, larger critical values are obtained. However, the degree to which the critical value is adjusted upward varies according to which test is used. Th , two post hoc tests discussed (e.g., Tukey, Scheffe) are conservative tests and the critical value is increased greatly in order to have more control over Type I errors (Huck, 2000).

REFERENCES

Elmore, P. , & Woehlke, P. (1998). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Goodwin, L. D. , & Goodwin, W. L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.

Hinkle, D. E., Wiersma, W., & Jurs, H. W. (1998). Applied statistics for the behavioral sciences. (4th ed.). Boston, MA: Houghton Mifflin Company

Huberty, C.J, & Morris, J.D. (1989). Multivariate analysis versus multiple univariate analysis. Psychological Bulletin, 105, 302-308.

Huck, S. W. (2000). Reading statistics and research. (3rd ed.). NY: Longman.

Love, G. (1988, November). Understanding experimentwise error probability. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Document Reproduction Service No. ED 304 451)

Thompson, B. (1994a). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. In B. Thompson (Ed.), Advances in social science methodology (Vol. 3, pp.3-27). Greenwich , CT: JAI Press.

Thompson, B. (1994b, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development (SWSRHD), Austin, TX. (ERIC Document Reproduction Service NO. ED 367 687)

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

Witte, R. S. (1985). Statistics (2nd ed.). New York: Holt, Rinehart and Winston.